International Journal of Advanced Engineering Research and Science (IJAERS)
Peer-Reviewed Journal
ISSN: 2349-6495(P) | 2456-1908(O)
Vol-10, Issue-12; Dec, 2023
Journal Home Page Available: https://ijaers.com/
Article DOI: https://dx.doi.org/10.22161/ijaers.1012.9

Convolutional Sparse Coding Multiple Instance Learning for Whole Slide Image Classification

Md Rony Molla, Ma Jian Fen

College of Computer Science and Technology (College of Data Science), Taiyuan University of Technology, China

Received: 25 Oct 2023,

Receive in revised form: 01 Dec 2023,

Accepted: 14 Dec 2023,

Available online: 21 Dec 2023

***Keywords*— Multiple Instance Learning, Weakly Supervised Learning, Whole Slide Imaging, Convolutional Sparse Coding**

Abstract— *Multiple Instance Learning (MIL) is commonly utilized in weakly supervised whole slide image (WSI) classification. MIL techniques typically involve a feature embedding step using a pretrained feature extractor, then an aggregator that aggregates the embedded instances into predictions. Current efforts aim to enhance these sections by refining feature embeddings through self-supervised pretraining and modeling correlations between instances. In this paper, we propose a convolutional sparsely coded MIL (CSCMIL) that utilizes convolutional sparse dictionary learning to simultaneously address these two aspects. Sparse dictionary learning consists of filters or kernels that are applied with convolutional operations and utilizes an overly comprehensive dictionary to represent instances as sparse linear combinations of atoms, thereby capturing their similarities. Straightforwardly built into existing MIL frameworks, the suggested CSC module has an affordable computation cost. Experiments on various datasets showed that the suggested CSC module improved performance by 3.85% in AUC and 4.50% in accuracy, equivalent to the SimCLR pretraining (4.21% and 4.98%) significantly of current MIL approaches.*

I. INTRODUCTION

The utilization of gigapixel resolution in digital whole slide imaging (WSIs) facilitates the comprehensive examination and analysis of the complete tissue sample within a singular image. Nevertheless, pathologists encounter substantial difficulties due to the magnitude and intricacy of the images [1]. Consequently, there is a growing need for automated workflows to facilitate WSI analysis. Because of this, deep learning-based methods have been increasingly used and developed in this sector [2,3,4,5,6,7,8,9]. The massive size of WSIs and the lack of annotations at the pixel level make deep learning approaches difficult to deploy [2]. To tackle these issues, approaches based on weakly-supervised multiple instance learning (MIL) have been suggested [6,7,8,9].

In the framework of MIL for WSI classification, each WSI is seen as a collection of non-overlapping patches that are extracted from the WSI slide. Each patch is considered as an unlabeled instance. The bag is classified as positive if at least one of the occurrences demonstrates the presence of disease, and negative otherwise. A commonly employed methodology for conducting MIL in WSIs involves a two-step process. Initially, the cropped patches undergo a process of conversion into feature embeddings by means of a fixed feature extractor. A fixed extractor is more desirable than a learned one since the computational cost of back-propagating with a large number of instances in a bag is prohibitively expensive. Next, a MIL aggregator is utilized to merge the embeddings of local instance features in order to get bag-level predictions. The potential sub-optimality of a two-stage learning scheme arises from the presence of noisy feature embeddings and imbalanced instances.

Specifically, the limited representation of positive instances within a positive bag can lead to the MIL aggregator learning an inaccurate mapping between embeddings and labels. Additionally, the limited supervisory signal poses a hindrance to the MIL aggregator's ability to capture correlations among instances [7, 8, 10].

Prior endeavors at MIL addressed these two obstacles individually. The initial category of approaches centered on enhancing the feature embeddings that were extracted through the use of self-supervised pretraining [7,10,11,12].

On the other hand, these techniques necessitate a large amount of data and an additional computationally intensive training phase. In order to eliminate negative instances, the second group of approaches concentrated on enhancing the MIL aggregator to better capture cross-instance correlations and imposing sparsity limitations on the local instance attention (e.g., picking k most significant instances) [13,14]. Better sparse instance feature embeddings that can model the invariance of the same type of biological tissues would also ease the duty of the MIL aggregator, therefore there is a strong relationship between these two types of approaches. Sparse feature embeddings, which are a low-dimensional version of the feature extractor's initial instance embeddings, are also advantageous to WSI representation since the high-dimensional WSI representation is located on a low-dimensional manifold, according to empirical evidence [10, 12].

Convolutional Sparse Coding (CSC) applies sparse representations to image or signal data and exploits local dependencies through convolutional processes. In signal processing and machine learning, sparse coding finds a sparse basis function representation of incoming data. We built an end-to-end learning-optimizable CSC module for convolutional sparse coding learning. Our new MIL framework, convolutional sparsely coded MIL (CSC-MIL), uses convolutional dictionary learning to improve initial feature embeddings. Traditional sparse dictionary learning algorithms are incompatible with deep neural networks and need considerable hyperparameter adjustment. Since it is complementary to current MIL frameworks, the proposed CSC module can be incorporated with them with reasonable extra processing. The experimental findings on different datasets and tasks proved that the suggested strategy helped state-of-the-art MIL methods perform better.

II. RELATED WORK

Methods for MIL can be broadly classified into two primary categories: instance-level MIL and bag-level MIL. In general, the instance-level approaches [15,16,17,18,19] include training a neural network to predict instance-level labels. These labels are assigned by propagating the bag-level label to each instance. The researchers combine the anticipated labels at the instance level in order to get the appropriate label at the bag level. However, as a consequence of the limited number of positive examples in a bag that are linked to a disease in whole slide images (WSIs), the negative cases within a positive bag are frequently incorrectly labeled. Despite multiple endeavors to refine the instance-level labels, empirical investigations have repeatedly demonstrated that instance-level approaches provide lower performance in comparison to their bag-level counterparts [8,20].

The bag-level multiple instance learning (MIL) approaches [6,7,8,9,20,13,21,22,23,14,12] employ a two-stage learning process. In this process, the methods initially transform the instances into a feature representation by utilizing a pretrained feature extractor. Subsequently, they employ MIL aggregation techniques to generate predictions at the bag-level. Previous investigations on bag-level MIL have predominantly concentrated on two main avenues. One potential option for improvement involves enhancing the MIL aggregator. The attention-based MIL [7,6] transformed non-parametric poolings like max/mean-pooling [20] into trainable ones using an attention mechanism. However, initial approaches examined each incident separately without considering similarities. Further research has addressed this restriction by using graph convolutional networks [14], non-local attention [7], transformers [8], and knowledge distillation [9]. The second approach involves enhancing feature embedding through self-supervised pretraining [7,10,11,12]. However, these approaches require ample data for task-specific training and are computationally costly.

The concept of employing sparse coding in network designs can be attributed to the research conducted by [24] who explored the application of unrolling sparse coding algorithms, such as ISTA, to acquire knowledge about the sparsifying dictionary. Several recent studies have investigated the utilization of deep networks using convolutional sparse coding layers for various tasks such as image denoising, picture restoration, and image classification with network normalization [25,26,27,28]. The efficacy of neural networks has primarily been established on datasets of limited or moderate sizes, particularly in the context of tasks such as picture categorization or production. In a recent study by [29], it was shown that convolutional sparse coding-inspired networks have achieved notable performance in image classification tasks using extensive picture datasets like ImageNet-1K.

Our work is similar to [10], which improved feature embedding using low-rank guided self-supervised pre-training and an attention-based MIL aggregator that utilizes

low-rank properties. However, it requires further self-supervised training and is customized for a certain MIL aggregator. Our approach may increase features and represent global instance similarities using a single module, making it easy to integrate into current MIL methods.

III. METHODOLOGY

To maintain the integrity of our analysis, we will focus on the specific case of bag-level binary MIL classification. The investigation aims to discover a correspondence between a collection of bags $x_1, x_2, \cdots, x_b$, and their respective labels $z_1, z_2, \cdots, z_b$, with $x_{i,j}$ is a positive integer greater than or equal to n instances $(x_{i,1}, x_{i,2}, \cdots, x_{i,n})$ and $\mathbb{R}^p$ is the dimension of each instance $(x_{i,1})$. The mathematical definition of the bag-level binary MIL classification is as follows:

$$y_i = \begin{cases} 0 & iff \sum_{j=1}^{n} y_{i,j} = 0 \\ 1 & otherwise \end{cases} \quad (1)$$

where $y_{i,j} \in < \{0,1\}$ represents the instance-level label of the *i*-th bag that is unknown, and n may differ between bags. A bag-level prediction is generated when a MIL aggregator aggregates the instance-level predictions contained within a bag.

$$\hat{y}_i = f_{cls}(\sigma(\phi_\omega(x_{i,1}), \phi_\omega(x_{i,2}), \cdots, \phi_\omega(x_{i,n})) \quad (2)$$

The function $f_{cls}(\cdot)$ represents a classifier at the bag level. The symbol ϕ_ω represents an embedding network that is parameterized by ω and operates at the instance level. The function σ is a permutation-invariant function. In this study, we examine four commonly used MIL pooling techniques, namely attention-based [6], non-local-based [7], transformer-based [8], and knowledge distillation-based [9].

3.1 SPARSELY CODED IN MIL

Empirical evidence supports the notion that the low-dimensional representation of instance embeddings significantly enhances the WSI representation [10, 12]. We assume the initial instance embeddings $\phi_\omega(x_i)$ can be represented in a low-dimensional space by a linear combination of $s << m$ atoms from an over-complete dictionary $D \in \mathbb{R}^{pxm}$, where m is the number of atoms in the dictionary. A classic Sparse Coding (SC) method for signals is to divide them into patches and solve for each

$$\min ||z||_0 s.t. x = Dz \quad (3)$$

3.2 CONVOLUTIONAL SPARSE CODING

The CSC model is derived from the classical SC model by exchanging the matrix with the convolutional operator.

$$x = \sum_{i=0}^{m-1} d_i * z_i \quad (4)$$

Where $x \in \mathbb{R}^{n_1 x n_2}$ is the input signal, $d_i \in \mathbb{R}^{kxk}$ a local convolution filter, and $z_i \in \mathbb{R}^{n_1 x n_2}$ a sparse feature map of the convolutional atom d_i the l_1 minimization problem for CSC formulated as

$$\arg\min_{d,z} \frac{1}{2} ||x - \sum_{i=0}^{m-1} d_i * z_i||_2^2 + \lambda \sum_{i=0}^{m-1} ||z_i||_1 \quad (5)$$

The CSC includes the entire input signal, unlike typical SC, which splits x into patches or segments. A learned atom of a certain edge orientation can globally represent all edges of that orientation in the image since the CSC model is spatially invariant.

Since convolutions are linear and the CSC model is a classical SC model, where D_{conv} is a concatenation of Toepltiz matrices, it can be interpreted to be a variant of classical SC. To format the aim in eq 5 replace the universal dictionary D with D_{conv} . Representing the CSC model as matrix multiplication is inefficient in memory and computation. Each element of x requires $n_1 x n_2 x m$ multiply and accumulate operations compared to convolution formation, which only requires s MACs (assuming $s << n_1, n_2$). ISTA iterations for CSC reads as:

$$z_{k+1} = S_{\lambda/L}(z_k + \frac{1}{L} d * (x - d * z_k)) \quad (6)$$

where $d \in R^{s \times s \times m}$ is an array of m s × s filters, $d * z = [flip(d_0) * x, \cdots, flip(d_{m-1}) * x]$ and $d * z = \sum_{i=0}^{m-1} d_i * z_i$. The $flip(d_i)$ operation flips the order of entries in d_i in both dimensions.

$$z_{k+1} = S_\theta(z_k + w_e * (x - w_d * z_k)) \quad (7)$$

the variables w_e, w_d, and θ are fully trainable and independent. We proposed the variable c to allow for numerous channels in the initial signal, such as color channels.

3.3 LEARNING THE OPTIMAL λ

The selection of the sparsity regularization strength λ is a crucial parameter in the ISTA with Acceleration. Tuning the value of λ is crucial for balancing sparsity and expressiveness in Convolutional Sparse Coding (CSC). However, within the framework of NA-MIL, the selection of the ideal λ value may differ between bags, posing a significant challenge for manual tuning. To achieve this objective, we framed the assessment of the optimal λ_i for each bag as a regression task. The parameter λ_i was represented as a feed-forward network (FFN) f_θ $(\phi_\omega (x_i))$. Here, θ represents the parameters of the network. The FFN in this study was composed of three fully-connected layers, and ReLU activation was performed on each layer. A mean pooling layer was incorporated following the fully connected feedforward neural network (FFN) to generate a

single numerical output for λ_i. The average pooling method was selected because the sparsity is similarly distributed across all target instance embeddings in a bag.

3.4 LEARNING THE OPTIMAL STEP SIZE

The choice of the step size has a crucial role in determining the convergence behavior of the ISTA with Acceleration (ISTA-ACC) algorithm. One common approach for learning the step size is line search, where the step size is determined dynamically at each iteration based on the properties of the objective function. Instead of utilizing a predetermined step size (α), it is suggested to employ a line search technique to choose the most favorable step size throughout each iteration. At every iteration k, conduct a line search in the direction of the negative gradient in order to get the optimal step size that minimizes the objective function. Two commonly used line search approaches in optimization algorithms are backtracking line search and quadratic interpolation.

Table. 1: Performance comparison on two classical MIL benchmark datasets

Method	Musk1	Musk2
mi-Net	0.886 ± 0.003	0.857 ± 0.002
MI-Net	0.887 ± 0.015	0.859 ± 0.012
Mi-Net with DS	0.894 ± 0.003	0.084 ± 0.002
Mi-Net with RC	0.967 ± 0.003	0.960 ± 0.002
ABMIL	0.892 ± 0.015	0.858 ± 0.012
ABMIL-Gated	0.900 ± 0.015	0.863 ± 0.012
GNN-MIL	0.917 ± 0.003	0.892 ± 0.002
DP-MINN	0.907 ± 0.003	0.926 ± 0.002
NLMIL	0.921 ± 0.003	0.910 ± 0.002
ANLMIL	0.912 ± 0.015	0.884 ± 0.012
DSMIL	0.932 ± 0.003	0.930 ± 0.002
ABMIL w/CSC	0.957 ± 0.015	0.957 ± 0.008
ABMIL-Gated w/CSC	0.969 ± 0.003	0.961 ± 0.002

IV. EXPERIMENTS

4.1 DATASETS

A series of tests were done on many datasets, encompassing two well-established MIL benchmarks, namely the MNIST-bags dataset [6], the CAMELYON16 dataset [30], and the Cancer Genome Atlas non-small cell lung cancer (TCGA-NSCLC) dataset. These experiments were carried out to assess and verify the efficacy of the proposed method.

Classical MIL benchmark datasets are MUSK1 and MUSK2. The MUSK1 and MUSK2 datasets estimate pharmacological effects based on molecular configurations. Each bag has several molecular conformations. The bag label is positive if at least one conformation has the intended pharmacological effect, and negative if none is effective [31].

The MNIST-bags dataset [6] contains random bags of grayscale handwritten digits from the MNIST dataset. As per [6], the digit of interest was '9', and any bag with at least one instance of it was considered affirmative. Ten incidences per bag were averaged, with a standard deviation of two. After adding the proposed SC module into attention-based MIL, we used this dataset for explanations.

The public WSI dataset CAMELYON16 detects metastatic breast cancer in lymph node tissue. The dataset contains 399 lymph node tissue WSIs (one corrupted sample was deleted), split into 270 training samples and 129 testing samples. Pathologists annotate each WSI with a binary label indicating metastatic cancer presence or absence in the lymph node tissue. Cancerous tissue areasof each WSI are also annotated in the dataset. After following the preprocessing steps in [7], we trimmed the WSIs into 224 × 224 non-overlapping patches. Approximately 3.37 million patches at ×20 magnification were produced, averaging 8451 per bag. The TCAGA-NSCLC dataset is utilized for the purpose of distinguishing between two sub-types of lung cancer, namely lung squamous cell carcinoma and lung adenocarcinoma. Following [7], we separated 1037 diagnostic WSIs into 744 training, 83 validation, and 210 testing sets. Following the same

Table.2: Evaluation of state-of-the-art approaches on CAMELYON16 and TCGA-NSCLC datasets. AUC and classification accuracy (%) were reported.

	Method		Camelyon16		TCGA-NSCLC	
			Accuracy	AUC	Accuracy	AUC
ResNet-18 ImageNet Pretrained	ABMIL-Gated		80.55	80.40	81.72	91.22
		+CSC	82.16	83.41	84.50	94.28
		Δ	+1.55	+4.13	+2.78	+2.06
	DSMIL		82.82	85.76	77.67	89.15
		+CSC	84.37	86.73	86.23	92.26
		Δ	+1.55	+0.97	+8.56	+3.11
	TransMIL		80.82	81.76	84.67	92.15
		+CSC	82.37	86.73	90.23	94.26
		Δ	+1.55	+4.97	+5.56	+2.11
	DTFD(maxS)		82.95	89.54	84.29	90.37
		+CSC	86.05	92.55	87.57	94.20
		Δ	+3.10	+3.01	+3.34	+3.83
ResNet-18 SimCLR Pretrained	ABMIL-Gated		82.05	82.05	85.72	90.22
		+CSC	85.16	86.41	88.50	93.28
		Δ	+3.11	+0.34	+2.78	+3.06
	DSMIL		86.82	85.76	86.67	93.15
		+CSC	88.37	87.73	90.23	95.26
		Δ	+3.11	+1.97	+3.56	+1.76
	TransMIL		86.82	85.76	86.67	93.15
		+CSC	88.37	87.73	90.23	95.26
		Δ	+3.11	+1.97	+3.56	+1.76
	DTFD(maxS)		82.95	89.54	84.29	90.37
		+CSC	86.05	92.55	87.57	94.20
		Δ	+3.10	+3.01	+3.34	+3.83

preprocessing as the CAMELYON16 dataset, about 10.30 million patches were retrieved at ×20 magnification. Each bag averaged 10355 patches.

4.2 BASELINES

We compared the proposed method against deep learning-based MIL methods, such as mi-Net and MI-Net [20], ABMIL and ABMIL-Gated [6], GNN-MIL [32], DP-MINN [33], and three non-local MIL pooling methods (NLMIL [34], ANLMIL [22], and DSMIL) on classical MIL benchmark datasets. For WSI classification, we investigated integrating the CSC module into four MIL aggregators: ABMIL with gated attention [6], DSMIL [7], TransMIL [8], and DTFD-MIL with MaxS[9].

4.3 EXPERIMENTAL SETTINGS

In this study, different experimental approaches were used for different datasets. We ran 10-fold cross-validation on conventional MIL datasets with five repetitions each experiment, focusing on classification accuracy. We tested the effectiveness of the suggested strategy by integrating the CSC module into the ABMIL framework utilizing two attention mechanisms: ABMIL w/ SC and ABMIL-Gated

Table.3: Proposed CSC module parameter selections the number of atoms in dictionary when L = 6

# Atoms(m)	# Params	FLOPS	AUC
m=4	59.53K	10.25K	86.61
m=8	94.03K	18.43K	88.31
m=16	189.13K	86.13K	90.45
m=32	561.65K	888.60K	90.75

w/ CSC. Additionally, 200 bags were used for training and 50 for testing in the MNIST-bags dataset. We used the training and testing partitions for the two WSI datasets. We tested features from two pretrained ResNet-18 models. Evaluation metrics included classification accuracy and AUC values.

This work trained all models using cross-entropy loss. All results in tables 1 and 2 had m = 16 atoms in a dictionary and L = 6 layers. The batch size was 1 for all tests. Models were trained on conventional MIL datasetsusing an Adam optimizer for 40 epochs, with initial learning of 1×10^{-4} and l2 weight decay of 5×10^{-4}. The initial learning rate was changed using a cosine annealing scheduler. We used the identical training settings on the MNIST-bags dataset but with a $5x10^{-4}$ initial learning rate and 1×10^{-4} weight decay. For the WSI classification tasks, we trained all four MIL aggregators for 200 epochs using default settings.

4.4 RESULTS

After integrating the proposed CSC module, ABMIL w/CSC and ABMIL-Gated w/CSC exceeded state-of-the-art classification accuracy approaches on all five MIL benchmark datasets (1). The ABMIL-Gated w/CSC outperformed the previous state-of-the-art accuracy by 2.5%, with 3.7% on MUSK1, 3.7% on MUSK2, and 3.1% on MUSK2. Furthermore, the accuracy of the ABMIL-Gated w/SC demonstrated the greatest stability, as evidenced by its average standard deviation of 0.0054.

Fig.1: Comparison of attention weight on positive bags on MNIST-bags ABMIL w/o CSC module and ABMIL w/ CSC module

The integration of the suggested CSC module consistently enhanced the performance of the four various types of MIL aggregators when combined with two distinct pre-training approaches, as shown in 2. This proves that the suggested CSC module's performance improvement is independent of the MIL aggregators and pre-training techniques used. Incorporating the suggested

CSC module into the CAMELYON16 dataset led to an average AUC improvement of 4.01% when using ImageNet pre-training and 2.60% when using SimCLR

Table.4: Proposed CSC module parameter selections the number of layer when atoms m = 16

#Layers (L)	FLOPS	AUC
L=2	59.53K	90.61
L=4	70.03K	91.31
L=6	86.13K	92.10
L=8	102.22K	91.27

pre-training. Applying the CSC module to two separate pre-trained feature embeddings also resulted in an average accuracy gain of 2.37% and 3.63%, respectively. We found that utilizing ImageNet pre-training increased AUC on the TCGA-NSCLC dataset by an average of 3.69%, whereas using SimCLR pre-training increased AUC by 2.28%. An accuracy enhancement of 4.01% and 6.63%, respectively, was observed when features derived from the two ResNet-18 models that were pre-trained differently were utilized. In addition, our results revealed that the enhancement in ImageNet pre-training (with an average AUC of 3.85% and an accuracy of 4.50%) was more substantial than that of SimCLR pre-training (with an AUC of 2.44% and an accuracy of 3.82%). This implies that the task of improving high-quality feature embedding, such as SimCLR pre-training, is more difficult compared to developing a low-quality feature embedding, such as ImageNet pre-training. Additionally, it is usually observed that superior feature embedding results in improved performance. Importantly, the suggested CSC module improved performance by 3.85% in AUC and 4.50% in accuracy, equivalent to the SimCLR pretraining (4.21% and 4.98%). However, in contrast to self-supervised pre-training, the proposed CSC module can be seamlessly integrated into existing MIL frameworks without requiring an additional training phase.

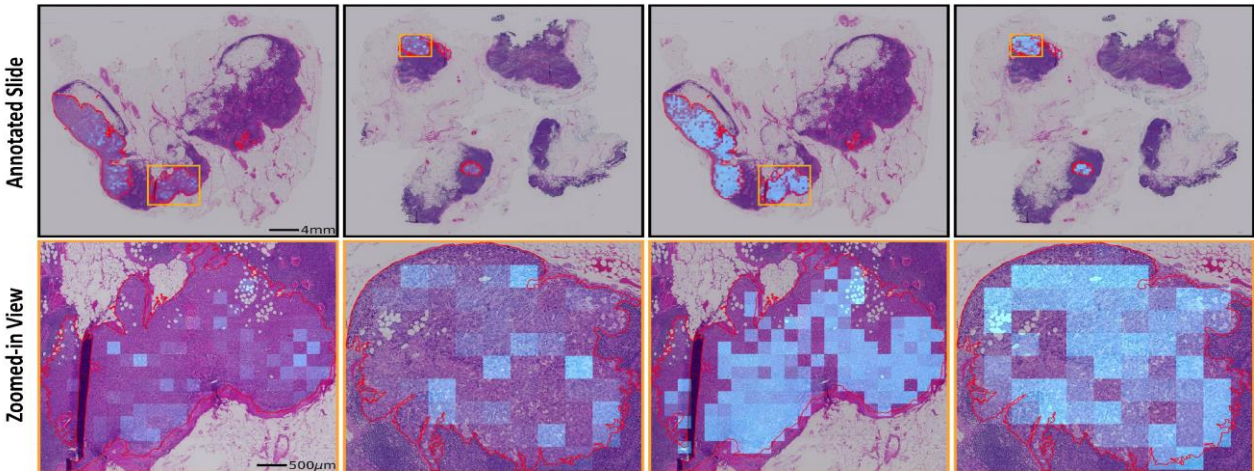

Fig. 2: The tumor localization in CAMELYON16 using the SimCLR pre-trained features: (a) map form ABMIL -Gated w/o CSC and (b) map form ABMIL-Gated w CSC

4.5 ABLATION STUDY

To examine the influence of hyperparameter selection (specifically, the number of layers L and the number of atoms in the dictionary m) on the performance of the CSC module, a sequence of ablation experiments were conducted. Using SimCLR pre-trained features on CAMELYON16, ABMIL-Gated performed ablation investigations. To investigate the influence of the quantity of atoms, we kept the number of layers constant at 6. Increasing the number of atoms resulted in a progressive performance improvement, as well as an increase in parameters and calculation (Table 3). We found that increasing the number of atoms from 16 to 32 improved performance by 0.32% in AUC but increased computation by tenfold. We fixed the number of atoms at 16 to test how layer count affected performance. Increasing the number of layers gradually made the AUC better, but it took more time to do the calculations (Table 4). A decrease in value was noted as L was raised from 2 to 4, potentially attributable to slight variations in the convergence trajectory of ISTA-ACC.

4.6 INTERPRETABILITY

In addition to enhancing the performance of numerous MIL aggregators, the proposed CSC module also improved their interpretability. The vanilla Attention-Based Multiple Instance Learning (ABMIL) model demonstrated an imbalanced distribution of attention scores (mean: 0.1780, standard deviation: 0.0748) when applied to the MNIST-bags dataset, specifically for the target digit '9'. (Fig. 1) This unequal distribution can be attributed to the model's lack of understanding regarding the relationships between instances within the dataset. Following the integration of the suggested CSC module, the attention scores for the ABMIL with CSC became more uniformly distributed (0.1995 ± 0.0341) (Fig. 1). Similar results were noted in the identification of the structure's location inside the designated area of interest in whole slide images (WSIs). The utilization of attention scores allowed for the assessment of the importance of each patch, hence offering valuable insights into critical morphological features that can inform clinical diagnosis. The vanilla ABMIL-Gated model had inadequate tumor localization performance on the CAMELYON16 dataset, as illustrated in Fig. 2, where it failed to accurately identify the majority of tumor patches.

Nevertheless, the incorporation of the CSC module greatly improved the localization accuracy of the ABMIL-Gated model, as depicted in Fig. 2, demonstrating a strong alignment with the annotated tumor shape. The results obtained from analyzing both the MNIST-bags dataset and the WSI dataset demonstrate that the proposed CSC module's coding of instance embeddings is capable of effectively capturing cross-instance similarities. This, in turn, results in improved localization performance.

V. CONCLUSION

Using Convolutional Sparse Coding learning, we presented a new MIL framework in this paper called CSCMIL. The method being suggested aims to improve both the embedding of instance features and the modeling of cross-instance similarities, all while minimizing the computational load. Significantly, empirical findings from numerous benchmarks spanning diverse tasks have demonstrated that the integration of the proposed CSC module in a plug-and-play fashion can enhance the performance of state-of-the-art MIL approaches. This

method has potential for drug effect prediction, diabetic retinopathy grading, and cancer detection and pathology analysis using histology.

ACKNOWLEDGEMENTS

We thank Ma Jian Fen and colleagues for helpful discussion, advises in high performance computing and editing..

REFERENCES

[1] He, L., Long, L.R., Antani, S.K., & Thoma, G.R. (2012). Histology image analysis for carcinoma detection and grading. Computer methods and programs in biomedicine, 107 3, 538-56 .

[2] Litjens, G.J., Sánchez, C.I., Timofeeva, N., Hermsen, M., Nagtegaal, I.D., Kovacs, I., Hulsbergen - van de Kaa, C., Bult, P., van Ginneken, B., & van der Laak, J.A. (2016). Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. Scientific Reports, 6.

[3] Zhou, X., Li, C., Rahaman, M.M., Yao, Y., Ai, S., Sun, C., Wang, Q., Zhang, Y., Li, M., Li, X., Jiang, T., Xue, D., Qi, S., & Teng, Y. (2020). A Comprehensive Review for Breast Histopathology Image Analysis Using Classical and Deep Neural Networks. IEEE Access, 8, 90931-90956.

[4] Kather, J.N., Pearson, A.T., Halama, N., Jäger, D., Krause, J., Loosen, S.H., Marx, A., Boor, P., Tacke, F., Neumann, U.P., Grabsch, H.I., Yoshikawa, T., Brenner, H., Chang-Claude, J., Hoffmeister, M., Trautwein, C., & Luedde, T. (2019). Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. Nature Medicine, 25, 1054 - 1056.

[5] Coudray, N., Ocampo, P.S., Sakellaropoulos, T., Narula, N., Snuderl, M., Fenyö, D., Moreira, A.L., Razavian, N., & Tsirigos, A. (2018). Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. Nature Medicine, 24, 1559-1567.

[6] Ilse, M., Tomczak, J.M., & Welling, M. (2018). Attention-based Deep Multiple Instance Learning. International Conference on Machine Learning.

[7] Li, B., Li, Y., & Eliceiri, K.W. (2020). Dual-stream Multiple Instance Learning Network for Whole Slide Image Classification with Self-supervised Contrastive Learning. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 14313-14323.

[8] Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., & Zhang, Y. (2021). TransMIL: Transformer based Correlated Multiple Instance Learning for Whole Slide Image Classication. Neural Information Processing Systems.

[9] Zhang, H., Meng, Y., Zhao, Y., Qiao, Y., Yang, X., Coupland, S.E., & Zheng, Y. (2022). DTFD-MIL: Double-Tier Feature Distillation Multiple Instance Learning for Histopathology Whole Slide Image Classification. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 18780-18790.

[10] Xiang, J., & Zhang, J. (2023). Exploring Low-Rank Property in Multiple Instance Learning for Whole Slide Image Classification. International Conference on Learning Representations.

[11] Lu, M.Y., Chen, R.J., Wang, J., Dillon, D., & Mahmood, F. (2019). Semi-Supervised Histology Classification using Deep Multiple Instance Learning and Contrastive Predictive Coding. ArXiv, abs/1910.10825.

[12] Li, H., Zhu, C., Zhang, Y., Sun, Y., Shui, Z., Kuang, W., Zheng, S., & Yang, L. (2023). Task-Specific Fine-Tuning via Variational Information Bottleneck for Weakly-Supervised Pathology Whole Slide Image Classification. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 7454-7463.

[13] Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., & Mahmood, F. (2020). Data-efficient and weakly supervised computational pathology on whole-slide images. Nature Biomedical Engineering, 5, 555 - 570.

[14] Zhao, Y., Yang, F., Fang, Y., Liu, H., Zhou, N., Zhang, J., Sun, J., Yang, S., Menze, B.H., Fan, X., & Yao, J. (2020). Predicting Lymph Node Metastasis Using Histopathological Images Based on Multiple Instance Learning With Deep Graph Convolution. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 4836-4845.

[15] Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A.P., Werneck Krauss Silva, V., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S., & Fuchs, T.J. (2019). Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nature Medicine, 25, 1301 - 1309.

[16] Feng, J., & Zhou, Z. (2017). Deep MIML Network. AAAI Conference on Artificial Intelligence.

[17] Hou, L., Samaras, D., Kurç, T.M., Gao, Y., Davis, J.E., & Saltz, J. (2015). Patch-Based Convolutional Neural Network for Whole Slide Tissue Image Classification. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2424-2433.

[18] Lerousseau, M., Vakalopoulou, M., Classe, M., Adam, J., Battistella, E., Carr'e, A., Estienne, T., Henry, T., Deutsch, É., & Paragios, N. (2020). Weakly supervised multiple instance learning histopathological tumor segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention.

[19] Xu, G., Song, Z., Sun, Z., Ku, C., Yang, Z., Liu, C., Wang, S., Ma, J., & Xu, W. (2019). CAMEL: A Weakly Supervised Learning Framework for Histopathology Image Segmentation. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 10681-10690.

[20] Wang, X., Yan, Y., Tang, P., Bai, X., & Liu, W. (2016). Revisiting multiple instance neural networks. ArXiv, abs/1610.02501.

[21] Sharma, Y., Shrivastava, A., Ehsan, L., Moskaluk, C.A., Syed, S., & Brown, D.E. (2021). Cluster-to-Conquer: A Framework for End-to-End Multi-Instance Learning for Whole Slide Image Classification. International Conference on Medical Imaging with Deep Learning.

[22] Zhu, Z., Xu, M., Bai, S., Huang, T., & Bai, X. (2019). Asymmetric Non-Local Neural Networks for Semantic Segmentation. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 593-602.
[23] Zhu, W., Lou, Q., Vang, Y.S., & Xie, X. (2016). Deep Multi-instance Networks with Sparse Label Assignment for Whole Mammogram Classification. bioRxiv.
[24] Gregor, K., & LeCun, Y. (2010). Learning Fast Approximations of Sparse Coding. International Conference on Machine Learning.
[25] Sreter, H., & Giryes, R. (2017). Learned Convolutional Sparse Coding. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2191-2195.
[26] Mohan, S., Kadkhodaie, Z., Simoncelli, E.P., & Fernandez-Granda, C. (2019). Robust and interpretable blind image denoising via bias-free convolutional neural networks. ArXiv, abs/1906.05478.
[27] Lecouat, B., Ponce, J., & Mairal, J. (2019). Fully Trainable and Interpretable Non-local Sparse Models for Image Restoration. European Conference on Computer Vision.
[28] Liu, S., Li, X., Zhai, Y., You, C., Zhu, Z., Fernandez-Granda, C., & Qu, Q. (2021). Convolutional Normalization: Improving Deep Convolutional Network Robustness and Training. Neural Information Processing Systems.
[29] Dai, X., Li, M., Zhai, P., Tong, S., Gao, X., Huang, S., Zhu, Z., You, C., & Ma, Y. (2022). Revisiting Sparse Convolutional Model for Visual Recognition. ArXiv, abs/2210.12945.
[30] Ehteshami Bejnordi, B., Veta, M., Johannes van Diest, P., van Ginneken, B., Karssemeijer, N., Litjens, G.J., van der Laak, J.A., Hermsen, M., Manson, Q.F., Balkenhol, M.C., Geessink, O.G., Stathonikos, N., van Dijk, M.C., Bult, P., Beca, F., Beck, A.H., Wang, D., Khosla, A., Gargeya, R., Irshad, H., Zhong, A., Dou, Q., Li, Q., Chen, H., Lin, H., Heng, P., Hass, C., Bruni, E., Wong, Q.K., Halici, U., Öner, M.Ü., Cetin-Atalay, R., Berseth, M., Khvatkov, V., Vylegzhanin, A., Kraus, O.Z., Shaban, M., Rajpoot, N.M., Awan, R., Sirinukunwattana, K., Qaiser, T., Tsang, Y., Tellez, D., Annuscheit, J., Hufnagl, P., Valkonen, M., Kartasalo, K., Latonen, L., Ruusuvuori, P., Liimatainen, K., Albarqouni, S., Mungal, B., George, A.A., Demirci, S., Navab, N., Watanabe, S., Seno, S., Takenaka, Y., Matsuda, H., Ahmady Phoulady, H., Kovalev, V.A., Kalinovsky, A., Liauchuk, V., Bueno, G., Fernández-Carrobles, M.D., Serrano, I., Deniz, O., Racoceanu, D., & Venâncio, R. (2017). Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. JAMA, 318, 2199–2210.
[31] Dietterich, T.G., Lathrop, R.H., & Lozano-Perez, T. (1997). Solving the Multiple Instance Problem with Axis-Parallel Rectangles. Artif. Intell., 89, 31-71.
[32] Tu, M., Huang, J., He, X., & Zhou, B. (2019). Multiple instance learning with graph neural networks. ArXiv, abs/1906.04881.
[33] Yan, Y., Wang, X., Guo, X., Fang, J., Liu, W., & Huang, J. (2018). Deep Multi-instance Learning with Dynamic Pooling. Asian Conference on Machine Learning.
[34] Wang, X., Girshick, R.B., Gupta, A.K., & He, K. (2017). Non-local Neural Networks. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 7794-7803.